Combining Wikipedia and WordNet for improving domain terms compilation

Jorge Vivaldi[a], Horacio Rodríguez[b], German Rigau[c]

[a] *Universitat Pompeu Fabra, Barcelona, Spain*
[b] *Polytechnical University of Catalonia, Barcelona, Spain*
[c] *University of the Basque Country, Donostia-San Sebastián, Spain*

Abstract

Domain terms are a useful mean for tuning both resources and NLP processors to domain specific tasks. This paper proposes an improved method for obtaining terms from potentially any domain using the Wikipedia graph structure as a knowledge source and the result of enriching WordNet with extended WordNet domains.

Keywords: term extraction, domain terminology, Wikipedia, Wordnet, Wordnet domains

1. Introduction and motivation

Sometimes NLP resources and tools claim to be domain independent, although its application to specific tasks uses to be restricted to specific domains. As the accuracy of such resources degrades heavily when applied in environments different from which they were built, a tuning to the new environment is needed.

The basic knowledge sources needed for performing this tuning are domain restricted corpora and terminological lexicons. Acquiring the latter is specially challenging and this is the goal of the work described here. Manual acquisition is costly and time consuming due to a low level of agreement among experts ([1]). We present here an approach for extracting terminology for a given domain using the Wikipedia (WP) and WordNet (WN). Although, in its conception, it is domain/language independent, we have applied it to the domains of Medicine and Economics and the English and Spanish languages.

2. Methodology

We obtain the terminology for a domain using the two WP graphs, WP_{PG} (pages) and WP_{CG} (categories), as knowledge sources. Our hypothesis is that page and category titles are candidates to domain terms (TC). From WP_{PG} and WP_{CG} we use the following types of edges: page $\rightarrow$ category and the inverse, category $\rightarrow$ category (super and sub-categories), with its inverse, page $\rightarrow$ page (input and output links from-to a page).

For getting the seed categories for starting we use the variants included in the synsets belonging to the domain. We use as domains those defined in WND [3]. WP and WN should be available for the language involved.

WND is a hierarchy of 169 domain labels which have been used to tag all WN synsets. These labels are organized into a shallow taxonomy. Information brought by domain labels is complementary to what is already in WN. A domain label can contain senses from different WN sub–hierarchies, include synsets of different syntactic categories, and subsume different senses of the same word.

However, the semi-automatic method used to develop WND produced errors and inconsistencies. [5] present a new robust graph-based method which propagates domain information through WN. They developed a new semantic resource called eXtended WordNet Domains (XWND) derived from WND and aligned to WN 3.0. For our process we use a normalized version of XWND.

The overall process, outlined in Figure 2, is iteratively applied to each pair ⟨domain, language⟩ independently. Let the pair ⟨*dc*, *lang*⟩ being $dc \in$ WND.

Firstly, we obtain the top categories in WP_{CG} ($\mathrm{categoryS}_0^{top}$) corresponding to *dc*. We get $\mathrm{categoryS}_0^{top}$ using with decreasing confidence WP_{CG}, WP_{PG}, page-category edges and interwiki edges. In most of the cases, e.g. Medicine, *dc* directly corresponds to a category in WP_{CG}, so $\mathrm{categoryS}_0^{top}$ consisted on just {'Medicine'}.

Secondly we extract from WN all the variants contained in all the synsets tagged in XWND with *dc*. We use WP_{CG} to analyze bottom-up such variants, resulting on $\mathrm{categoryS}_0^{dc}$. Categories in $\mathrm{categoryS}_0^{dc}$ are scored taking into account (i) their XWND score and (ii) whether they have as ancestors the elements in $\mathrm{categoryS}_0^{top}$ and the distance to tops. Using a threshold we obtain an initial set of in domain categories $\mathrm{categoryS}_0^{dc.inic+}$ and a complementary set of off-domain categories $\mathrm{categoryS}_0^{dc.inic-}$. For Medicine the sizes of these sets were respectively 253 and 2,263.

Thirdly, $\mathrm{categoryS}_0^{dc.inic+}$ are top down expanded traversing WP_{CG} following the subcategory links, avoiding cycles, filtering out neutral categories and categories placed in WP_{CG} above the domain tops and discarding the expanded categories and descendents when belonging to $\mathrm{categoryS}_0^{dc.inic-}$. In this way the final $\mathrm{categoryS}_0^{dc+}$ is obtained. For Medicine the size of this set was of 1,924 categories.

Fourthly, we get from $\mathrm{categoryS}_0^{dc+}$ the first set of candidate categories. We have applied 5 different selection methods (m) to this task, differing on whether only $\mathrm{categoryS}_0^{top}$ or $\mathrm{categoryS}_0^{dc+}$ is used for computing the distances to the tops and on the use of a complementary category classifier learn from ⟨Medicine, English⟩ manually evaluated data. $\mathrm{categoryS}_0^{m}$ result from this step.

Fifthly, the initial set of pages, pageS_0^{m}, is built. From each category in $\mathrm{categoryS}_0^{m}$ the set of belonging pages, following category-page links, is collected. Each category is scored according to the scores of the pages it contains and each page is scored according both to the set of categories it belongs to and to the sets of pages linked to it. Three thresholding mechanisms are used: (i) *Microstrict* (accept a category if the number of member pages with positive score is greater than those with negative score), (ii) *Microloose* (similarly with greater or equal test), and (iii) *Macro* (instead of using the page scores we use the scores of the categories of the pages).

Then, in step 6, we iteratively explore each category repeating the same process again. The set of well scored pages and the set of well scored categories reinforce each other. Less

scored categories and pages are removed at each iteration, so the global precision of the sets is expected to grow at a cost of a draw in recall. A combination function is used for computing the global score of each page and category from their constituent scores. The process is iterated until convergence, leading in iteration i to categoryS$_i^m$, pageS$_i^m$. These sets are collected for all the iterations and selection methods.

In step 7 a final filtering is performed for selecting from all the categoryS$_i^m$ and pageS$_i^m$ corresponding to all the iterations and selection methods in step 4 the one with best F1.

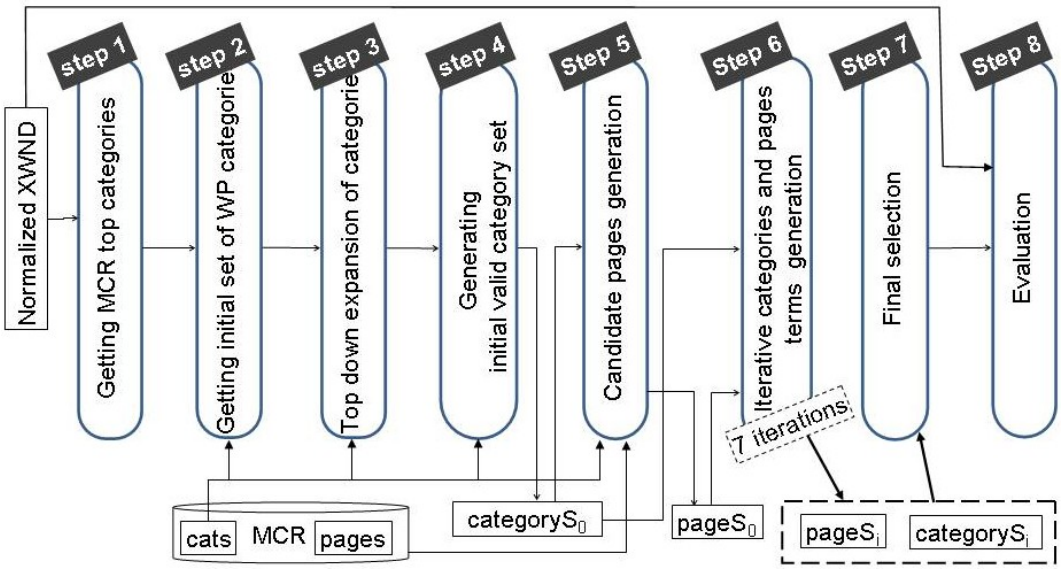


Figure 1: Methodology

3. Evaluation

In the process of creating a resource, evaluation plays an important role. The first drawback is the nonexistence of gold standards to evaluate against as well as objective comparison methods. Evaluating a terminology is a difficult task ([1]) due to: *a*) the difficulty in doing it through human specialists (therefore it becomes a subjective task), *b*) the lack/incompleteness of electronic reference resources and *c*) disagreement among them (specialists and/or reference resources).

Looking to minimize the above mentioned problems, we set up two different scenarios for evaluating the resources obtained with our system:

- Due to the lack of reliable references for most of the domains, we planned a first scenario doing a partial evaluation, restricted to terms occurring both in WN and WP; that can be applied to any domain/language.

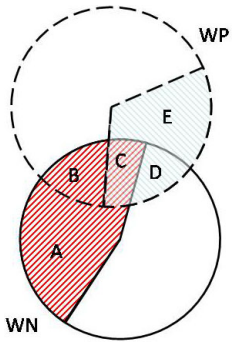
- Instead, for those few domains where an external reliable reference is available a full evaluation scenario was foreseen. This is the case of Medicine for which we use SNOMED, a well known medical term repository.

Our guess is that accuracy results can be extrapolated to terms not occurring in WN and to domains lacking external references. First consider the content of Figure 2 that shows the basic sets of TCs that must be considered in doing this evaluation. This evaluation is based on the subset C as they are the only terms that are tagged in WN as belonging to the domain and at the same time should take part of the set of pages/categories found in WP. Precision and recall can be easily calculated as indicated in formulae 1 and 2.

For this purpose we use two baseline systems for comparison. The first, *Magnini-baseline*, consists on, giving *dc*, collecting all the synsets assigned to it, and considering as TCs all the variants related to them. This approach has the obvious limitation of reducing coverage to the variants contained in WN; also it is rather crude because no score is attached to TCs, despite their degree of polisemy or domainhood. The second, *NG-baseline*, is based on [2]. It maps WP pages with WN synsets. Our baseline is built collecting all the synsets corresponding to *dc* and from them all the WP pages aligned with them.

Figure 2: Terms indirect evaluation.

A: WN domain variants not found in WP
B: WN domain variants found in WP but not recovered by our system
C: WN domain variants found in WP
D: WN variants belonging to the domain according the WP but not according WN
E: WP pages/categs belonging to the domain but not found in WN
A+B+C: WN variants for a given domain
C+D+E: WP pages/categories recovered



$$Precision = |C|/(|C| + |D|) \tag{1}$$

$$Recall = |C|/(|B| + |C|) \tag{2}$$

4. Experiments

The first step, as shown in Figure 1, was obtaining the WN variants associated to the *dc* Medicine. Table 1 shows all the *dc* related to Medicine and for each the number of variants found in MCR and those found in WP and SNOMED.

As mentioned in section 2, we experimented 5 ways to filter out terms that do not belong to the domain. The results obtained in applying such methods to Medicine are summarised in Table 3. For the best case ($m = 0$) we collected 22,311 TC (pages and categories). From this set of TC, only 4,614 (20.7 %) exist in WN. Consequently the partial evaluation was done on this set.

Table 1: Number of variants across the medical subdomains

Subdomain	MCR	MCR+WP	%	SNOMED
medicine	4380	2952	67.40	2664
anatomy	3808	2572	67.54	1691
physiology	1022	696	68.1	477
psychology	892	612	68.61	175
pharmacy	167	120	71.86	73
genetics	85	64	75.29	24
psychiatry	57	41	71,93	24
surgery	39	31	79.49	23
dentistry	26	22	84.62	8
radiology	20	16	80	13
Total	10496	7126	67.89	5172

Table 2: Results of the partial evaluation over Medicine (excluding subdomains)

system	count	precision	recall	F1
ours	1016	0.65	0.81	0.72
Magnini_baseline	1257	0.53	1.00	0.69
NG_baseline	324	1.00	0.26	0.41

4.1. Evaluation issues

As shown in Table 3, none of the proposed methods improve the base method (0). The method that uses multiples tops has become the second better method as its F-score is closed to the base method although the number of terms discovered is much lower (-20%).

The results of our first (partial) evaluation are presented in Table 2. Due to the definitions of baselines, Magnini's gets a perfect recall while NG's gets a perfect precision. Our system ranks in between. Our F1 outperforms slightly Magnini's and largely NG's.

The use of SNOMED allows a more serious evaluation. As shown in Table 3. (with tag *) recall is consistently improved at a cost of small drop in precision. F1 reaches 54.9. These results are more realistic because they are evaluated against a reference lexicon. Nevertheless there are some inadequacies in using this repository. See for example the following terms included in both MCR and WP as belonging to the medical domain but not present in SNOMED:

- *first-aid kit*
- *bloodletting*
- *medical report*
- *maxillary*
- *abductor*: but SNOMED includes 51 similar terms like: *Abductor pollicis muscle*, *Abductor of wrist joint*, etc.

Table 3: Precision, recall and F-score for all defined methods for Medicine/English (* against SNOMED)

Method		0	1	2	3	4
		CategoryS_0^{top}	CategoryS_0^{top} + classifier	CategoryS_0^{dc+}	CategoryS_0^{dc+} + classifier	Intersect
Category	num	1016	572	793	283	366
	Prec./Rec.	70.92/3.62	81.64/3.16	78.69/3.47	84.97/1.97	92.19/2.68
	F-score	6.89	6.08	6.65	3.85	5.21
	Prec./Rec.*	63.5/5.67	72.66/4.93	71.13/5.48	76.47/3.10	71.13/5.48
Page (strict)	num	14639	9066	11278	5339	7286
	Prec./Rec.	69.11/34.03	80.76/25.98	74.41/32.38	80.68/15.54	83.12/23.32
	F-score	45.61	39.32	45.13	26.06	36.42
	Prec./Rec.*	52.80/45.52	80.76/25.98	57.02/43.46	60.09/20.27	61.43/30.18
Page (loose)	num	21295	11472	16923	8490	9567
	Prec./Rec.	64.77/42.60	77.32/32.96	69.95/41.01	79.15/25.50	80.05/30.42
	F-score	51.40	46.22	51,71	38.57	44.08
	Prec./Rec.*	49.40/56.89	77.32/32.96	53.43/54.85	60.50/34.13	60.53/40.28

Table 4: Details of the iteration that reaches the best F-score

		Categories	Pages	All terms
MCR	OK	239	2815	3034
	KO	98	1531	1580
	F1	6.89	51.40	49.56
	precision	70.92	64.77	65.02
	recall	3.62	42.59	40.98
SNOMED	precision	63.50	49.40	49.98
	recall	5.67	56.88	54.77

- *lachrymation*: but SNOMED includes the orthographical variant *lacrimation*.

Most of the above sequences seem to be terminological but some are discarded because their low terminological value or just missed. Instead, SNOMED include some other sequences that do not seems to be terminological like: *blue devil*, *voice box* or *puffing*. Results are presented in Table 4. Looking at the terms not found in SNOMED, most of those that correspond to diseases (*aneisokonia*, *vitiligo*, ...) have an ICD-10 code [1]. In the rest, a few of them (like *yawn*) cannot be considered as terminological units.

[1]ICD-10 is a medical classification list by the World Health Organization

Another problem come from the fact that WN includes several variants in a synset; although often only a few of them are included in SNOMED. For example the synset "00831191-n" contains the following variants: *breathing*, *ventilation*, *external respiration*, *respiration*. Only the first two are included in SNOMED, this fact lows the recall.

5. Conclusions and future work

In this paper we presented a new approach for obtaining the terminology of a domain using the category and page structures of WP and the XWND resource in a language/domain independent way. This approach has been successfully applied to the English medical domain showing a clear improving in relation previous approaches. As foreseen, the results evaluation is difficult, mainly due to inadequacies in the reference repository. Also the encyclopedic character of WP conditioned the list of new terms obtained.

The current definition of domain (a set of *dc*'s) could be problematic when considering subdomains, domains with no clear borders (like medicine, biology and chemistry) or interdisciplinary domains (like law, environment or information science). This will be a topic for future research/improvement.

The final version of this paper will include a detailed state of the art, will enlarge methodology discussion as well as will include complete details for two domain (Medicine and Economics) and two languages (English and Spanish).

6. Acknowledgements

References

[1] Vivaldi J., Rodríguez H. *Using Wikipedia for term extraction in the biomedical domain: first experience.* In Procesamiento del Lenguaje Natural 45, p. 251-254 (2010).

[2] E. Niemann, Gurevych I. *The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet.* In: Proceedings of the 9th International Conference on Computational Semantics, p. 205-214 (2011).

[3] Bentivogli, L., Forner, P., Magnini, B., Pianta, E.: *Revising WordNet Domains hierarchy: Semantics, coverage, and balancing.* In: Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources (2004)

[4] Zesch T., Müller C., Gurevych I. *Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary.* In LREC 2008: Proceedings of the Conference on Language Resources and Evaluation, p. 1646-1652, Marrakech (2008).

[5] Gonzalez-Agirre A., Castillo M. and Rigau G. *A proposal for improving WordNet Domains.* 8th international conference on Language Resources and Evaluation (LREC'12). Istambul, Turkey (2012).